IDA DOCUMENT D-715

THE CURRENT STATUS OF RESEARCH AND DEVELOPMENT ON SELECTION AND CLASSIFICATION OF ENLISTED PERSONNEL

Jesse Orlansky
Institute for Defense Analyses

Frances Grafton
Army Research Institute

Clessen J. Martin
Office of the Chief of Naval Operations

William Alley
Air Force Human Resources Laboratory

Bruce Bloxom
Defense Manpower Data Center

June 1990

Prepared for
Office of the Under Secretary of Defense for Acquisition
(Research and Advanced Technology)

INSTITUTE FOR DEFENSE ANALYSES
1801 N. Beauregard Street, Alexandria, Virginia 22311-1772

IDA Log No. HQ 90-35308

DEFINITIONS

IDA publishes the following documents to report the results of its work.

Reports

Reports are the most authoritative and most carefully considered products IDA publishes. They normally embody results of major projects which (a) have a direct bearing on decisions affecting major programs, (b) address issues of significant concern to the Executive Branch, the Congress and/or the public, or (c) address issues that have significant economic implications. IDA Reports are reviewed by outside panels of experts to ensure their high quality and relevance to the problems studied, and they are released by the President of IDA.

Group Reports

Group Reports record the findings and results of IDA established working groups and panels composed of senior individuals addressing major issues which otherwise would be the subject of an IDA Report. IDA Group Reports are reviewed by the senior individuals responsible for the project and others as selected by IDA to ensure their high quality and relevance to the problems studied, and are released by the President of IDA.

Papers

Papers, also authoritative and carefully considered products of IDA, address studies that are narrower in scope than those covered in Reports. IDA Papers are reviewed to ensure that they meet the high standards expected of refereed papers in professional journals or formal Agency reports.

Documents

IDA Documents are used for the convenience of the sponsors or the analysts (a) to record substantive work done in quick reaction studies, (b) to record the proceedings of conferences and meetings, (c) to make available preliminary and tentative results of analyses, (d) to record data developed in the course of an investigation, or (e) to forward information that is essentially unanalyzed and unevaluated. The review of IDA Documents is suited to their content and intended use.

The work reported in this document was conducted under contract MDA 903 89 C 0003 for the Department of Defense. The publication of this IDA document does not indicate endorsement by the Department of Defense, nor should the contents be construed as reflecting the official position of that Agency.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 1990	Final--May 1989 to January 1990

4. TITLE AND SUBTITLE

The Current Status of Research and Development on Selection and Classification of Enlisted Personnel

5. FUNDING NUMBERS

C - MDA 903 89 C 0003

T - T-D2-435

6. AUTHOR(S)

Jesse Orlansky, Frances Grafton, Clessen J. Martin, William Alley, Bruce Bloxom

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
1801 N. Beauregard St.
Alexandria, VA 22311-1772

8. PERFORMING ORGANIZATION REPORT NUMBER

IDA Document D-715

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

OUSD(A)/R&AT
The Pentagon, Room 3D129
Washington, DC 20301

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Research and development programs conducted by the military services and defense agencies on the selection and classification of enlisted personnel are discussed. Particular attention is given to (1) the use of computer-based testing procedures to measure performance capabilities not otherwise observable; (2) extensive efforts to validate the results of selection instruments with on-the-job performance data; and (3) ways to improve our ability to use test results and data processing procedures that better match people to available jobs. Suggestions are made for future research and development.

14. SUBJECT TERMS

Selection, classification. military personnel, research, development, testing procedures, computerized testing, aptitude testing, Armed Forces Qualification Test

15. NUMBER OF PAGES

72

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

IDA DOCUMENT D-715

THE CURRENT STATUS OF RESEARCH AND DEVELOPMENT ON SELECTION AND CLASSIFICATION OF ENLISTED PERSONNEL

Jesse Orlansky
Institute for Defense Analyses

Frances Grafton
Army Research Institute

Clessen J. Martin
Office of the Chief of Naval Operations

William Alley
Air Force Human Resources Laboratory

Bruce Bloxom
Defense Manpower Data Center

June 1990

INSTITUTE FOR DEFENSE ANALYSES
Contract MDA 903 89 C 0003
Task T-D2-435

ACKNOWLEDGMENTS

This work was prepared under Task Order T-D2-435, Cost Effectiveness of Human Factors and Training Technology, under the technical cognizance of Earl A. Alluisi, then the Assistant for Training and Personnel Technology, Deputy Director Defense Research and Engineering (Research and Advanced Technology), Office of the Under Secretary of Defense (Acquisition). We would like to acknowledge the support of Dr. Alluisi during this study, and also that of Wayne S. Sellman, Office of the Assistant Secretary of Defense (Force Management and Personnel).

In addition, the authors acknowledge the helpful suggestions of Franklin L. Moses, of the Army Research Institute, and Bart Kuhn, of the Office of the Chief of Naval Operations, as well as the contributions of many representatives of the military services and of the defense agencies who participated in a comprehensive review of research and development on enlisted personnel selection and classification.

ABSTRACT

Research and development programs conducted by the military services and defense agencies on the selection and classification of enlisted personnel are discussed. Particular attention is given to (1) the use of computer-based testing procedures to measure performance capabilities not otherwise observable; (2) extensive efforts to validate the results of selection instruments with on-the-job performance data; and (3) ways to improve our ability to use test results and data processing procedures that better match people to available jobs. Suggestions are made for future research and development.

CONTENTS

ABBREVIATIONS

AA	Aptitude Area
ABLE	Assessment of Background and Life Experience
ACAP	Accelerated CAT-ASVAB Program
AFQT	Armed Forces Qualification Test
AIT	Advanced Individual Training
ARI	Army Research Institute
ASD	Assistant Secretary of Defense
ASVAB	Armed Services Vocational Aptitude Battery
CAST	Computerized Adaptive Screening Test
CAT	Computer Adaptive Testing
CG	Commanding General
CL	Clerical
CONUS	Continental U.S.
CV	Concurrent Validation
DON	Department of the Navy
EPAS	Enlisted Personnel Allocation System
EST	Enlistment Screening Test
IRT	Item Response Theory
JOIN	Joint Optical Information Network
LAMP	Learning Abilities Measurement Program
LRDB	Longitudinal Research Data Base
LV	Longitudinal Validation

MEPS	Military Entrance Processing Station
METS	Mobile Examining Team Site
MM	Mechanical Maintenance
MOS	Military Occupational Specialty
MRA&L	Manpower, Reserve Affairs and Logistics
NCO	Non-Commissioned Officer
NPRDC	Navy Personnel Research and Development Center
OSUT	One Station Unit Training
PACE	Processing and Classification of Enlistees
PJM	Person-Job Match
P&P	Paper-and-Pencil
PROMIS	Procurement Management Information System
R&D	Research and Development
SC	Surveillance/Communication
SME	Subject Matter Expert
SQT	Skill Qualification Test
SynVal	Synthetic Validation
TOW	Tube-launched, Optically-tracked, Wire-guided
TRADOC	Training and Doctrine Command
TTP	Time-To-Proficiency
USAREUR	U. S. Army Europe
USMEPCOM	U.S. Military Entrance Processing Command

EXECUTIVE SUMMARY

Jesse Orlansky

Institute for Defense Analyses

Alexandria, Virginia

EXECUTIVE SUMMARY

The military services used selection tests during World War I and, since then, have contributed many significant improvements to their utility. In 1979 it was found that the norm used in the then-current selection test battery for military enlistment was not correct. This was a result of a switch in reference groups from the mobilization population of 1944 to the youth population of 1980. Congress directed the Department of Defense to link (i.e., validate) entry test scores to on-the-job performance rather than to performance during initial training. Rapid developments in testing theory and in computer technology were available to support this effort.

The services undertook many new programs to respond to the problems that had been identified and to the Congressional guidance. A Topical Area Review of Testing R&D and Planned Applications to Enlisted Personnel Selection and Classification was held on December 8-9, 1988. Representatives of the following organizations described their work on selection and classification:

Army Research Institute

Office of the Chief of Naval Operations

Navy Personnel Research and Development Center

Air Force Systems Command

Air Force Human Resources Laboratory

Defense Manpower Data Center.

Proceedings of that review have been published (Orlansky, Alluisi, and Sellman, 1989).

The purpose of this paper is to:

(1) Summarize the findings and views presented at the review,

(2) Discuss their adequacy and relevance for use in establishing policy on, and implementation of, procedures for screening, selection and classification of enlisted personnel,

(3) Make recommendations for R&D that is needed to fill gaps or take advantage of new developments, and

(4) Suggest a plan and schedule for completing the R&D and for implementing the results of this program.

The validation of the Armed Forces Vocational Aptitude Battery (ASVAB) can be improved significantly with new on-the-job performance measures (Project A) and with new second-tour performance data (Project B). Test administration will be improved by computer-based adaptive screening and computer-based adaptive aptitude test batteries. Further improvements for selection, classification and assignment should be sought by using computer-based procedures to measure capabilities not previously measurable in the cognitive, spatial, perceptual and psychomotor domains. It also appears possible to link selection and classification standards with a military hierarchical job structure that would significantly improve the person-job match.

The current status of R&D on the selection and classification of enlisted personnel, being performed by each military service, is described in the main body of this report. The order of presentation of that material is an arbitrary one.

I. CURRENT STATUS OF R&D ON SELECTION AND CLASSIFICATION OF ENLISTED PERSONNEL

A. PROJECT A (1982-1989): IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED PERSONNEL.

1. Armed Services Vocational Aptitude Battery (ASVAB)

Improved ASVAB Aptitude Area composite scores for the Clerical and Surveillance/Communication areas were introduced in October 1984. These changes improved the accuracy and fairness of predicting job performance for minorities. It is estimated that recommended changes in composite scores for approximately 50 Military Occupational Specialties (MOS) will produce savings of $25 million per year.

2. The New Enlisted Personnel Allocation System (EPAS)

This system will more efficiently match qualified Army applicants to occupations for which they are best qualified. The potential savings to the Army of using ASVAB, the improved composites, and EPAS to meet requirements is estimated to exceed $480 million per year.

3. Spatial and Computerized Psychomotor and Perceptual Tests

These tests, developed in Project A, have improved the validity of classification algorithms (e.g., multiple r of 0.76 for tank gunnery, measured in simulators.) Similar results were found firing the tube-launched, optically-tracked, wire-guided (TOW) missile simulator.

4. Assessment of Background and Life Experience (ABLE)

This multiple-choice, non-cognitive test provides scores on temperament, personal history, and adaptability. It improves the prediction of attrition, and in-service discipline problems, and its converse, prediction of high performance, non-commissioned officers.

5. Longitudinal Research Data Base (LRDB)

Empirical information collected during Project A is anticipated to have high value for examining issues related to accession policy, standards for enlistment and reenlistment, attrition, school training, and field performance.

B. COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)

The CAST, a computerized adaptive test, is administered at recruiting stations in order to estimate the likely performance of potential recruits on the Armed Forces Qualification Test (AFQT). It can be completed in 15 minutes, compared to an equivalent paper-and-pencil test that takes 45 minutes. Against the AFQT, it demonstrated validities of about 0.80, in three trials. It is used as a screening device to reduce the number of ASVABs that would otherwise have to be given to recruits. The use of CAST has reduced the number of ASVAB tests given to Army prospects from over one million per year in the early 1980s to about 400,000 per year currently while the number of accessions per year has remained about the same.

C. BUILDING AND RETAINING THE CAREER FORCE (PROJECT B)

The objectives of this program are to develop selection and classification procedures that will optimize the performance of soldiers in their second tour. This effort will use the longitudinal data from Project A and also determine the validity of ASVAB, new predictors, training performance, and first tour performance in predicting future, including second tour, performance.

D. SYNTHETIC VALIDATION PROJECT

This is an exploratory effort to use the Longitudinal Research Data Base to estimate the predictive validity of tests for new Military Occupational Specialties without performing empirical validity evaluations.

E. COMPUTER ADAPTIVE TESTING-ARMED SERVICES VOCATIONAL APTITUDE BATTERY (CAT-ASVAB)

The Computer Adaptive Testing-Armed Services Vocational Aptitude Battery (CAT-ASVAB) Program was initiated in 1979 as a Joint-Service effort to develop and deploy an automated testing system to replace the paper-and-pencil version of ASVAB. The Navy was designated as Executive Agent for development of the CAT system. The

original plan was modified in 1985 and renewed as the Accelerated CAT-ASVAB Program (ACAP) in order to deploy off-the-shelf equipment. This program involves six studies:

1. Pre-test (November 1980)

Led to revisions in the instructions to the test and reducing its reading grade level from the eighth to the sixth grade level.

2. Medium of Administration

This study examined the influence that method of administering the ASVAB (computer-based or paper-and-pencil) would have on calibrating test items. Analyses are now under way.

3. Cross-Correlation Study

This study compared the precision of the computer-based and paper-and-pencil versions of the ASVAB. Analyses are under way.

4. Preliminary Operational Check

This research showed that test data can be transferred accurately between the Data Handling Computer at a Military Entrance Processing Station (MEPS) and the ACAP System. Future tests involve sending the information by electronic means to U.S. Military Entrance Processing Command (USMEPCOM) Headquarters.

5. Score Equating Development Study

This effort is designed to equate the CAT and a paper-and-pencil version of ASVAB. Testing has been completed at six sites and data are being analyzed.

6. Score Equating Verification Study

This effort is designed to evaluate the effect of personal motivation on item calibration and item equating for both versions of the ASVAB. The schedule for collecting test data is July 1990 to September 1991.

F. CAT-ASVAB COST-BENEFIT ANALYSIS

Four ways of administering the CAT-ASVAB were compared to the current paper-and-pencil procedure:

1. Use of centralized and MEPS-only testing facilities.
2. Use of 283 mobile testing vehicles, and 50 more high volume testing sites, in addition to the present 70 MEPSs.
3. Increasing the number of high volume sites to 273, in addition to the present 70 MEPSs.
4. Computerized Adaptive Screening Test (CAST) administered by recruiters; complete CAT-ASVAB testing only at the present 70 MEPSs.

The utility benefits of CAT-ASVAB were based on an assumed increase of 0.002 in predictive validity; this value comes from a simulation study. The life cycle costs of all CAT-ASVAB concepts were higher than the paper-and-pencil ASVAB system. The latter costs about $15 million a year per 1 million tests. Since the investment needed for the computerized versions ranged from $15 to $40 million per million tests, it was concluded that larger increases in validity were needed to make any new CAT-ASVAB testing procedure cost-effective. Current efforts in the CAT-ASVAB program are directed towards increasing the validity, and thereby the cost-effectiveness, of computerized ASVAB testing. Increases in the validity of selection tests could have a large impact both on cost savings and in fleet readiness. For example, an increase in the validity of a test battery of 3 percent over the current ASVAB is estimated to be worth about $83 million per year in performance improvement in the Navy. An increase of one percent in the average Shop Practices Test scores of Boiler Technicians on two-screw, 1200 psi ships would lower Casualty Report downtime by an average of 138 hours per month per ship.

II. RECOMMENDATIONS FOR R&D THAT IS NEEDED TO FILL GAPS OR TO TAKE ADVANTAGE OF NEW DEVELOPMENTS.

On the basis of the significant progress that has occurred in this area, R&D is needed on the following aspects of testing and its applications to the selection and classification of enlisted personnel:

1. New Tests/Measures/Assessment Strategies

The search for new ways to measure individual differences that have implications for military performance should be concentrated in the following broad areas:

a. Cognitive Measures

In the cognitive domain, we need to be concerned with extending the ASVAB into areas that presently are not measured or not measured very well. R&D activities need to proceed along two fronts, each of which benefits from the other: (a) basic theory and model-building coupled with (b) empirical investigations of what can be shown to relate (beyond ASVAB) to subsequent performance. Areas showing the most promise take advantage of recent advances in cognitive psychology. In place of "fixed abilities" (i.e., "g," or verbal and quantitative abilities), we should be assessing more fundamental component processes--e.g., speed and accuracy of cognitive processing, capacities for short (working) and long-term memory and retention, prior declarative and procedural knowledge, etc.

The most innovative approaches will capitalize on the added precision and storage capacity of the microcomputer as a testing device. Unlike paper-and-pencil tests, administration of test materials via microcomputer brings not only additional control to the test situation, but also extends the domain of what can be measured effectively (i.e., response latency, multiple tasking, audio/visual stimulus cues, and multiple response modes).

b. Spatial, Perceptual, and Psychomotor Measures

The ASVAB is especially weak in its ability to measure performance in the spatial and perceptual-psychomotor domains. Here again, there is an opportunity to exploit the enormous potential of the microcomputer to present spatial stimuli that are impossible to replicate with paper-and-pencil tests. The same holds true for extending our selection measures to include perceptual-psychomotor abilities that have shown special promise in situations where performance demands are especially high. These would include such specialty occupations as air traffic controller and fighter pilot.

c. Non-Cognitive Measures

The whole area of non-cognitive measures (biographical data forms, vocational interest inventories, personnel security assessments, etc.) remains untapped. The research community has come to realize that selection procedures should not focus exclusively on increasing expected competencies, but instead should look at a broader perspective that considers the length of expected service (attrition/retention) and performance qualities that are not well predicted with cognitive test batteries, e.g., leadership/management, personnel integrity, and organizational commitment.

d. Training Performance Measures

Traditional reliance on training performance measures as the sole criterion reference should decrease in the future. For the past ten years, each of the Services has mounted large efforts to construct on-the-job performance measures (hands-on as well as other procedures). We now have a research data base that is unprecedented in the research literature. It needs to be probed to determine (a) how present tests relate to job criteria; (b) how much new tests can add to the prediction of job criteria; and (c) how alternative and potentially less expensive surrogate measures can be used in lieu of the more expensive hands-on measures.

e. Effect of Individual Performance on Group Effectiveness

As the next major step in this enterprise, research needs to concentrate on the effect of individual performance on crew, group, and unit success. Unless it can be shown definitively how individual differences in ability affect performance of military tasks, e.g., sortie generation capacity or tank crew effectiveness, we will have missed an important opportunity to show how entry standards influence combat effectiveness.

2. New Job/Task Analytic Procedures

Progress on this front will key on the extent that traditional job analysis techniques--which in the military are vertically oriented toward the requirements of single specialties in isolation--can be made to look and operate across specialties. One of the key factors, it turns out, that determines how many people are needed to perform a given amount of work, how much aptitude they must possess, and how much they need to be trained, is the specialty structure defined by the personnel system. Considering that some new systems technology is revolutionary rather than evolutionary, the Manpower Personnel and Training needs of the future are presently ill-defined. Research in job and task analytic procedures is needed to better understand present and future requirements for those who maintain and operate our weapons systems.

What is needed is a structured approach to job analysis that permits analysis across specialties. Such a system could better inform our attempts to structure specialties based on common entry and training requirements. It would also allow a more flexible response to changes in operational requirements (i.e., central versus dispersed basing), permit more economical approaches to training for "common skills," and eventually provide the basis for estimating requirements for jobs that do not yet exist but that can be described in detail.

3. New Selection and Classification Systems/Specifications

Future research on selection and classification systems can best be described in the context of the present multitiered system in which at least four different levels of processing are distinguished (see Fig. 1).

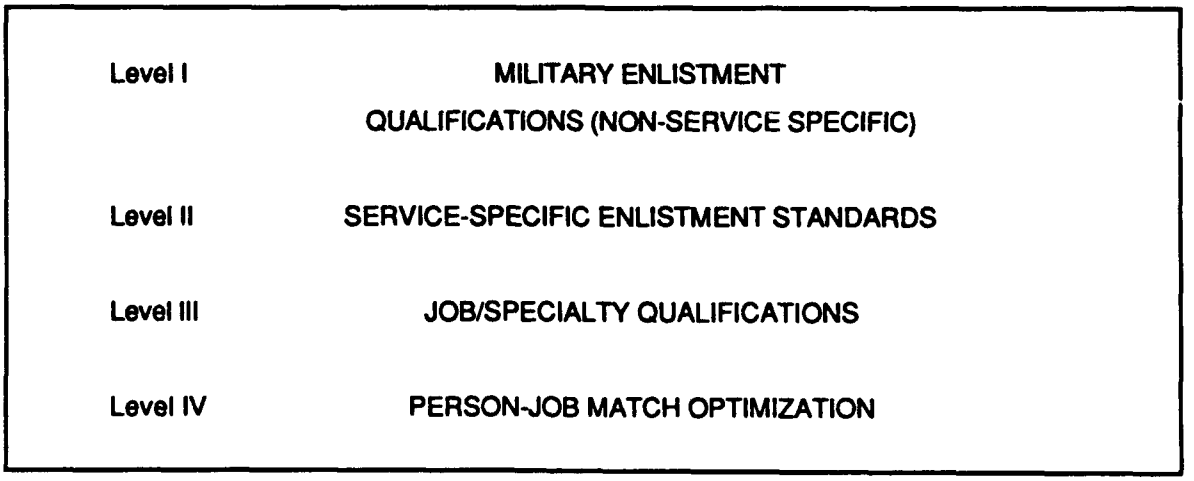

Figure 1. Multitiered selection and classification system

Levels I, II, and III operate on a "qualified-not qualified" basis by specifying minimum test scores that must be attained for entry. Level IV (Person-Job Match) processing assumes that all previous minimum qualifications have been met and assigns people to jobs such that the total system performance payoff is maximized.

The technology available for setting minimum qualifying standards is fairly rudimentary at present. Basically, the problem is one of setting a cutoff value high enough to ensure training and job success but not so high as to unduly restrict the supply of applicants. This is usually done on the basis of expert judgment and consensus after relevant empirical findings have been reviewed. These might include relationships between various test scores and training success, current training attrition rates, the occupational learning difficulty of the job, and the recruiting environment among others.

"Analytical" approaches to setting minimum standards, as exemplified by the RAND cost/performance trade-off model or the Air Force's Time-To-Proficiency (TTP) model, have recently been developed but these, as a general class, suffer from being either overly simplistic or much too narrow in scope to provide meaningful results. At present, the analytical approach represents more of an idealized future target than a near-term-solution to the problem. Work needs to proceed on these models to make them more effective representations of real-world situations. Procedures need to be developed for measuring productive capacity in individuals and in functional units (e.g., work centers, squadrons). Simulation capabilities based on analytical models should be extended from the present focus on first-term capacities to second and subsequent terms to account for the fact that over time, the value of individual workers proceeds from primarily technical activities in the early part of one's career to primarily supervisory/management activities in the latter part.

Alternative "eclectic" approaches have already been adopted by the Services' attempt to weigh judgmentally the myriad of personnel, job, and "management" factors that need to be brought to bear on decisions about who should be selected/classified into what jobs. From the testing side, these approaches require tests and composites that have demonstrated validity and the ability to differentiate between alternative assignments. From the job requirements side, they require that we be able to characterize jobs in terms of learning difficulty and in terms of the relevant aptitudes necessary for success. In neither of these areas are we anywhere near an optimal specification. The job clusters presently in use by the Services (e.g., Mechanical, General, Administrative and Electronics) evolved over a number of years and are in serious need of restructuring. The same is true of the

associated test composites upon which entry into the jobs within clusters is based. And within clusters, the systems employed for determining minimum cut scores are far from rigorous.

These systems also operate under the presumption that appropriate cognizance has been given to balancing the needs for *efficient* job fill with those of providing for *effective* job fill. The optimization (Level IV) systems in place today (i.e., PROMIS, PACE, EPAMs, etc.) are strong on "conceptual development" but weak in terms of follow-on "evaluation" or trade-off capability. That is, we need better techniques to track the overall payoffs over time, to decompose the total payoff into component benefits, and to determine in advance the effects of alternative reconfigurations on the system prior to actually making changes.

4. Technology Base Efforts

a. Multiple Criteria

In many real-world situations, there are multiple criteria around which a selection system should be designed. These may be simultaneous, as in selection, to maximize training performance, as measured by academic grade and instructor evaluations, or they could be sequential--selection to improve training outcomes, job performance and retention. Relationships between the predictors and criteria can be mutually complementary, mutually antagonistic, or independent. Similarly, some management preference usually exists for satisfying one or another of the criteria at the possible expense of the others. Research is needed to develop analytical strategies for coping with this situation that may involve both empirical and judgmental components.

b. Test Equity and Bias

Current methods for detecting and removing bias from formal selection systems are in many ways inadequate. R&D needs to explore the most effective and efficient means for determining the presence or absence of bias and for taking positive steps to reduce or eliminate it.

c. Acquisition and Decay of Abilities

Developmental theories relating the acquisition and decay of test-relevant abilities over the age groups 18-55 years have not kept pace with the Services' need to understand this important phenomena. As a result, educational effects have not been adequately

considered in developing test norms, policies for retesting are little more than ad hoc, and we have little understanding how overall capabilities to acquire knowledge or perform effectively change over time.

d. Automated Test Development

New approaches are needed to streamline the cumbersome and time-consuming task of developing follow-on test forms. Automated item generation is in its infancy--much more could be done to develop expert systems in this area. Once items are banked, procedures would be needed to extract items for parallel forms such that content, item difficulty, item discrimination, etc., were matched to the reference forms.

e. Dishonest Test Scores

With large financial incentives at stake, there is much to be gained by employing possibly illegal means to increase test scores. Detection systems at present are useful mainly in the aggregate, i.e., identifying problems among groups of examinees in a particular geographic region. Additional work in this area could provide a much more accurate estimate of whether cheating has occurred with regard to specific individuals and by how much.

f. Population Changes in Ability Levels/Patterns

Currently, there are no established procedures for tracking changes in aptitudes/ abilities among the eligible applicant population. Methods and procedures are needed to chart and project changes as they occur and as they might be expected to occur over the next 10-20 years. Such data could be very relevant to military planners who must trade off technological solutions with the demands created for the future operators and maintainers.

REFERENCES--EXECUTIVE SUMMARY

Orlansky, Jesse, Alluisi, Earl A., and Sellman, Wayne S. *Testing R&D and planned applications to enlisted personnel selection and classification: Proceedings of a topical area review. December 8-9, 1988.* IDA Document D-573, January 1989. Institute for Defense Analyses, Alexandria, Virginia.

I. IMPROVING THE SELECTION, CLASSIFICATION AND UTILIZATION OF ARMY ENLISTED PERSONNEL

Frances Grafton

Army Research Institute

Alexandria, Virginia

CONTENTS

I. IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED PERSONNEL

A. INTRODUCTION

Beginning in 1979, military testing has been the focus of increasing attention, from both the Department of Defense and Congress. The reasons for this interest are varied and include:

1. The discovery in 1979 that the then-operational selection test battery for military enlistment was misnormed.
2. The switch in reference groups for military tests from the 1944 mobilization population (all men, including officers, serving under arms during that year) to the 1980 youth population (all 17- to 23-year-olds in 1980 including both males and females).
3. A Congressional mandate to link entry test scores directly to on-the-job performance rather than to traditional training performance.
4. Intense interest in combining computer technology with emerging developments in testing theory.

As an integrated response to the various concerns, the U.S. Army Research Institute designed Project A, "Improving the Selection, Classification, and Utilization of Army Enlisted Personnel," a multi-year, multi-million-dollar research effort. Project A had many objectives including (1) continued validation (ASVAB), (2) development and validation of new cognitive and noncognitive measures that might expand the coverage of the predictor domain beyond that of the ASVAB in predicting total soldier performance, and (3) development of a comprehensive set of performance criteria including paper-and-pencil tests of school knowledge and job knowledge, hands-on tests of job performance, and behaviorally anchored rating scales.

Project A was initiated in 1982 and will be completed in 1989. While Project A focused primarily on Army enlisted personnel in their first tour of service, a second research program, "Building and Retaining the Career Force," is scheduled to begin in 1990. This effort will focus on the Army's career force, those soldiers who reenlist and

remain beyond one tour of duty. Soldiers who comprise the career force provide not only leadership and continuity to the Army as an institution, but are crucial in ensuring that the Army meets its battlefield mission.

Another accomplishment by the Army in military testing during the last decade has been the development and implementation of the Computer Adaptive Screening Test (CAST). CAST is used in recruiting stations to determine whether a prospect is likely to score well enough on the ASVAB to be considered for enlistment. Prior to implementation of CAST, all prospects' eligibility to enlist had to be determined by transporting them to a testing site to take the three-and-a-half-hour ASVAB. The use of CAST has reduced the number of ASVAB tests given to Army prospects from over one million per year in the early 1980s to approximately 400,000 per year currently, while the number of accessions per year has remained about the same.

Recognizing that large-scale efforts like Project A will not be feasible for all Army jobs, an exploratory research effort, "The Army Synthetic Validation Project," is currently under way. This program is designed to develop procedures for (1) deriving prediction equations for MOSs using primarily the extensive data gathered on the relatively small sample of MOSs in Project A, and (2) setting selection standards linked to job performance.

The remainder of this paper will describe more fully these efforts as well as some of the resulting products that have been implemented either operationally or in testbeds. Future plans in the area of military testing will also be discussed.

B. PROJECT A

In Project A, the research objectives are to:

1. Develop new, comprehensive measures to cover the job performance domain in the Army. These measures include both Army-wide and MOS-specific rating scales, written school and job knowledge tests, and direct hands-on measures on MOS-specific task proficiency.
2. Validate ASVAB against both existing and new performance measures.
3. Develop and validate new selection and classification measures against existing and new performance measures.
4. Validate intermediate criteria (such as training performance), as predictors of later criteria, such as first-tour job performance, so that better informed decisions on reenlistment can be made.

5. Examine the validity and utility of alternative procedures for making operational selection and classification decisions in the Army.

The Project A research design consists of three main stages of data collection and analysis in an iterative progression of development, testing, evaluation, and further refinement of the selection/classification instruments (predictors) and measures of job performance (criteria). In the first stage, individual-level data from FY 1981-82 were examined to explore relationships between soldiers' ASVAB scores and their later performance in training and on first-tour Skill Qualification Tests (SQT).

Prior to the second stage of the data collection, 19 Military Occupational Specialties (MOS) were selected as a representative sample of the Army's 250+ entry-level jobs. These MOSs were selected based on (1) a clustering of MOSs based on similarity of job content, (2) the fact that these MOSs are representative of Army jobs in general and account for over 45 percent of the Army's accessions, and (3) the recommendations of General Officers representing major Army Commands at that time. Nine of the 19 MOSs were then selected to have both paper-and-pencil and hands-on tests of job performance in addition to the school knowledge tests and ratings administered in all the MOSs. The nine fully tested MOSs in Project A are: Infantryman (11B), Cannon Crewman (13B), Tank Crewman (19E), Radio Operator (31C), Light Wheel Vehicle Mechanic (63B), Administrative Specialist (71L), Motor Transport Operator (88M), Medical Care Specialist (91A), and Military Police (95B).

The second stage of the Project A research, the Concurrent Validation (CV), was conducted in FY 1985. During the CV, over 9400 soldiers in the 19 MOSs were administered the new Project A predictor tests, which included measures of spatial abilities, temperament and vocational interest, as well as computerized tests of perceptual and psychomotor skills. These tests were designed to expand the predictor domain in terms of individual characteristics and attributes that might be important for selection in predicting more aspects of performance. The new predictors were intended to supplement those cognitive abilities already assessed by ASVAB.

Concurrently, the CV soldiers, who had been in the Army 18 to 27 months, were administered a comprehensive set of job performance measures including supervisory and peer ratings, written school and job knowledge tests, and MOS-specific hands-on task proficiency measures.

One of the major scientific contributions of the Project A research to date is a comprehensive modeling of the job performance domain. The criterion development efforts in the project were driven by the idea that job performance is multidimensional and that performance is best measured through a variety of methods. Analyses of the criterion data collected during the CV resulted in five empirically derived dimensions that represent overall Army job performance. The five Project A job performance dimensions are:

1. MOS-Technical Knowledge and Skill: The proficiency with which a soldier performs the technical tasks that are critical and central to his/her MOS.
2. General Soldiering Proficiency: How well a soldier executes common soldiering tasks such as first aid and land navigation.
3. Effort and Leadership: The degree to which a soldier exerts effort over the full range of job tasks, perseveres under adverse or dangerous conditions, and demonstrates leadership and support toward peers.
4. Personal Discipline: The degree to which a soldier adheres to Army regulations, exhibits self-control and does not create disciplinary problems.
5. Physical Fitness and Military Bearing: The degree to which a soldier maintains an appropriate military appearance and remains in good physical condition.

The first two performance dimensions involve a soldier's ability to perform the technical requirements of his/her job and are often referred to as "can do" factors. The remaining three performance dimensions are more attitudinal or motivational in nature and are often called "will do" factors.

The third stage of the Project A research, known as the Longitudinal Validation (LV), started in 1986 with the administration of the battery of new Project A predictor tests to approximately 55,000 new recruits in 21 MOS. At the end of his/her Advanced Individual Training (AIT) or One Station Unit Training (OSUT), each soldier was administered the school knowledge test for his/her MOS, and a set of rating scales was collected from supervisors and peers. Approximately 11,000 of these soldiers were followed into their first-tour field assignment and have been administrered a set of MOS-specific and Army-wide job performance measures during FY 1988 and FY 1989. In addition, about 1,000 of the soldiers from the CV sample who reenlisted have been assessed on a battery of performance measures appropriate for second-tour including rating scales, written job knowledge tests, hands-on measures, and leadership measures.

C. RESULTS AND PRODUCTS

During the execution of the research plan, Project A made many scientific contributions to the field of Industrial and Organizational Psychology and, so far, has delivered four major products to the Army.

1. ASVAB

One product which resulted from the Concurrent Validation involves improvements in the computation and use of ASVAB Aptitude Area (AA) Composites. These composites are used to determine eligibility for training in Army jobs. Beginning with the operational implementation of ASVAB 11/12/13/14 in October 1984, the formulae for the Clerical (CL) and the Surveillance/Communications (SC) composites were changed. These changes resulted in both better accuracy and improved fairness in the prediction of job performance for minorities in several MOSs. In addition, a change in the computation of the Mechanical Maintenance (MM) composite and recommended changes in required aptitude areas for approximately 50 MOSs are scheduled for implementation in the near future. Annual savings from the changes in composites have been estimated at $25 million.

The Enlisted Personnel Allocation System (EPAS) is a new assignment system that will more efficiently match qualified Army applicants to jobs for which they are best qualified, maximizing performance and minimizing attrition. Thus, with significant improvements in classification composites (from Project A) and in assignment (from EPAS), the Army can put the "right" person in the "right" job at the "right" time. The potential savings to the Army from using ASVAB (including the improved composites) and the Enlisted Personnel Allocation System to fill accession requirements is estimated to be more than $480 million annually.

Results from the Project A CV have demonstrated conclusively that ASVAB is an excellent test of the "can do" or more technical task-based requirements of Army jobs. The mean validity coefficient between scores on ASVAB and Core Technical performance is 0.57 across the 19 MOSs in the Concurrent Validation. Since ASVAB was designed as a general cognitive test, it is not expected to be a good measure of more specialized spatial abilities, psychomotor skills, motivation, interests, or leadership. As a means of supplementing ASVAB, the new Project A tests measure these characteristics very well.

2. Spatial, Psychomotor, and Perceptual Tests

Currently, the Army Research Institute is supporting implementation of Project A's spatial and computerized psychomotor and perceptual tests in USAREUR and at training posts in CONUS.

These implementations were inspired by ARI selection research on tank and anti-tank gunners. In 1986, Project A's new tests had a multiple R of 0.76 against accuracy on high-tech simulators of tank gunnery for 95 Armor officers at Fort Knox. Cognitive ability, which was tested as well, did not contribute to the prediction. In 1987, 300+ new recruits at Fort Benning took the same battery before training in anti-tank gunnery. Several of the new tests strongly predicted accuracy in firing the TOW (Tube-launched, Optically tracked, Wire-guided) missile simulator.

Based on these results, in December 1987, CG TRADOC ordered implementation of Project A tests at four posts that train on high-tech weapons systems. ARI installed 1- and 2-hand tracking, maze, and mental rotation tests at Forts Knox, Benning, and Bliss in February 1988.

At the Infantry and Armor sites, the earlier results have been confirmed. For 1,065 TOW students to date, 41 percent of those passing a cut score on the predictors qualified as gunners in the minimum time, as opposed to 24 percent of those not tested on the psychomotor/spatial predictors and 12 percent of those scoring below the cut score. Those scoring above the cut score qualified higher as well as faster: 0.48 percent attained the upper two levels of accuracy as contrasted with 36 percent of those not tested and 31 percent of those below the cut score. In terms of validities, accuracy was predicted at 0.37 by the Project A tests, 0.29 by ASVAB GT, and 0.38 by the two together. In mid-1988, these training data were confirmed in live fire at Fort Benning: for 60 students firing one live TOW each at a moving target at 6,000 ft, P(hit) was 0.85 for those who met the cut on the Project A tests and 0.73 for those who did not.

At Fort Knox, Armor recruits (N = 500+) scoring in the upper third on the new predictors early in training later had gunnery hit rates 16 percent higher than those in the lower third, exactly repeating the results of the 1986 research. The new predictors correlated 0.54 with speed/accuracy, compared with a 0.34 validity for the ASVAB GT composite. Combined, the validity was 0.55, reconfirming the importance of spatial and psychomotor skills in predicting gunnery performance. At both posts, the new results were so positive that the test scores were incorporated into the decisionmaking process.

Validities at Fort Bliss were not significant, and testing there has been suspended. In contrast to tank and anti-tank weapons, which rely heavily on tracking, the air defense systems tested use fire-and-forget weapons against fleeting targets. Thus, these systems rely more on skills like vigilance and target identification.

Completing the implementation requests from CG TRADOC, the Field Artillery School, Fort Sill, is starting to administer a broader battery of the new tests to recruits in meteorology, surveying, radar range-finding, and artillery spotting (MOS). Validation will be against performance at the end of training.

Currently, the active forces are transitioning to the Bradley Fighting Vehicle. In USAREUR, V Corps has elected to use a battery of spatial and psychomotor tests to inform their selection of Bradley gunners in the 3rd Armor Division and the 8th Infantry Division.

A need for improved selection has been found also at the Special Forces School, where attrition, due primarily to failures in land navigation, is costly. In a new pilot project, ARI has installed three spatial tests at Fort Bragg to identify good land navigators.

In review, early positive results are being replicated in the implementations; spatial/psychomotor abilities strongly predict differences in gunnery performance; and utilization is spreading. Budget reductions, however, threaten the survival of the implementation in the training base. In the future, the tests could have the greatest impact if they were administered before enlistment. In that case, the Army Reserve and National Guard could be served as well, and initial person-job matching could be strengthened in many MOSs by adding spatial and/or psychomotor abilities to the profile of aptitudes.

3. Assessment of Background and Life Experience (ABLE)

The ABLE is a 30-minute, multiple-choice, non-cognitive test designed to measure temperament, personal history, and adaptability. In Project A, ABLE was shown to improve significantly the prediction of the motivational components of performance. Scores on the Adjustment scale of ABLE are strongly related to 12-month attrition. Recruits who have very low Adjustment scores have attrition rates that are two to three times higher than soldiers with high Adjustment scores.

The Dependability scale of ABLE predicts in-service disciplinary problems. For example, soldiers having low Dependability scores receive significantly more Articles 15

than those with high scores. Conversely, soldiers with high Dependability scores are more often viewed as having potential of becoming high performing NCOs.

In sum, the ABLE shows promise to augment the Services' capability to identify the most qualified applicants in all AFQT categories and educational levels by predicting the "will do" components of performance. Based on these encouraging results, the ABLE is included in a joint-Service adaptability screening instrument that will become operational in 1990.

4. The Longitudinal Research Data Base

The fourth, and perhaps most enduring product from the Project A research effort, is the Longitudinal Research Data Base (LRDB). The LRDB is a permanent storehouse of empirical information on Project A, unparalleled in its richness for addressing recurring Army concerns. These data are invaluable to the Army in the areas of accession policy, setting standards for enlistment and reenlistment, predicting attrition, linking school training to field performance, and linking characteristics at entry into the military to performance in first and second tours of duty.

D. BUILDING AND RETAINING THE CAREER FORCE

At the end of the Project A Longitudinal Validation data collections in FY 1989, the first phase of the Army's research program to build a new selection and classification system, based on job performance, will be completed. The second phase, which will include measures of critical tasks for soldiers in their second tour, will be initiated in FY 1990.

The research objectives planned for Building and Retaining the Career Force are to:

1. Develop a set of measures for selecting and classifying enlisted personnel in order to optimize second tour soldier performance without sacrificing first tour performance.
2. Analyze longitudinal data from Project A and other sources to develop information, procedures, and recommendations for implementing new selection and classification measures in building the Army's career force.
3. Determine the validity of ASVAB, new predictors, training performance, and first tour performance in predicting future performance (including second tour).

Results generated from both Project A and Building and Retaining the Career Force will be useful to Army policymakers in setting enlistment standards and to Army training proponents in setting quality distribution goals during years of a declining manpower pool. The outcomes will also be helpful to DoD in providing (1) more tools for conducting selection and classification testing, and (2) scientific data to use in responding to Congress and the general public on questions concerning military personnel testing.

E. COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)

The Computerized Adaptive Screening Test (CAST) is a tool for recruiters to estimate the likely performance of a prospect on the Armed Forces Qualification Test (AFQT) of the ASVAB. As the name suggests, CAST is a computer adaptive test, and it is administered in recruiting stations. The computer selects succeeding items based on responses to previous items. Because item difficulty is matched to examinee ability, maximum information is obtained from each item. Thus, a short CAST can provide as much information as a longer paper-and-pencil test.

CAST was preceded by the Enlistment Screening Test (EST). The EST was a good predictor of individuals' scores on the AFWT, and second, the EST has to be hand scored. The implementation of JOIN (Joint Optical Information Network) provided computers for recruiting stations, and computer-administered testing became feasible. Whereas the EST, a paper-and-pencil test, takes 45 minutes, CAST can be completed in 15 minutes.

CAST was initially developed by the Navy Personnel Research and Development Center (NPRDC) in a cooperative arrangement with ARI. A field test on 364 Army applicants showed a validity of 0.85 using AFQT as the criterion. This was an encouraging finding, comparable to the validity of 0.83 for EST. Two later cross-validations on larger samples yielded comparable results (i.e., validities of 0.80 and 0.79).

The initial implementation of CAST received a generally positive reception from recruiters and was successful in predicting operational AFQT. However, some improvements were necessary, leading to recent refinements in CAST. The number of items in CAST was increased from approximately 300 to over 500. This increase was needed to reduce the likelihood that any particular item would be compromised as a result of overuse. In addition, a better format for reporting results was needed. The original display predicted a specific AFQT score. What the recruiter and prospect really wanted to know, however, was the likelihood that the prospect would score in certain critical ranges of the AFQT. Accordingly, the final display was redesigned to meet that need. The current

version of CAST also has been calibrated to the new AFQT introduced operationally in January 1989.

F. SYNTHETIC VALIDATION PROJECT

The Army's Synthetic Validation (SynVal) Project is an exploratory research effort currently under way. The concept of synthetic validity was first introduced by Lawshe in 1952 as a middle ground between situational validity, which required costly validity analyses (empirical validation) for each job, and generalized validity, which assumes the validity of a test across similar jobs. The project A research, with its comprehensive set of new predictor tests (non-cognitive, spatial, and psychomotor) in addition to the existing ASVAB and extensive job performance measures, afforded a unique opportunity to examine SynVal as a cost-effective alternative to empirical validity studies for a large number of MOS.

This project has two main goals:

1. Develop procedures for identifying job performance prediction equations for new MOSs, low-fill MOSs, and for other MOSs when it is impractical or too costly to derive empirical prediction equations.
2. Develop procedures for establishing cut scores on predictor tests that are linked to job performance and that identify both minimally qualified and highly qualified recruits.

The research design has three phases, each with a data collection. The initial phase concentrates on MOSs in the Project A sample that had the full complement of criterion measures; the second phase adds MOSs which had reduced criterion testing; and the final phase will apply the SynVal procedures to at least one MOS that was not included in Project A. Comparisons will be made among job-component models and among standard-setting procedures.

Three job-component models have been developed and tested: (1) Job Behaviors Model, (2) Attribute Model, and (3) Job Tasks Model. The results of phase one have been extremely encouraging. Army subject matter experts (SMEs) were able to reliably use the three models to identify job components (in terms of importance or estimated validity) for each of the three MOS. The resulting job description (component) profiles differed systematically across MOS. Furthermore, the average "synthesized" validity (across the three phase one jobs and models) compared favorably with the average empirical validity

(0.50 versus 0.67, respectively). However, given the exploratory nature of these findings, it would be premature to exclude any model or weighting scheme.

With respect to the second goal, an extensive literature review found little research on procedures for setting cut scores that are linked to job performance. The project assumes that requirements for different levels of job performance must be determined first. Then, cut scores could be set on the predictor (selection) measure(s) to optimize the likelihood of obtaining individuals who would meet the specific performance requirements.

Four performance levels have been defined objectively: Unacceptable, Marginal, Acceptable, and Outstanding. The three standard setting approaches developed and tested in phase one were: (1) Soldier-based, (2) Task-based, and (3) Critical Incident-based. Although mixed results have been obtained, the existing literature on standard setting also consistently indicates differences in obtained standards with different techniques. The critical-incident approach resulted in the most lenient standards for minimal performance followed by the soldier-based approach with the task-based approach being the most stringent. However, the task-based approach resulted in more variation (less agreement) among the judges while the critical-incident approach showed the greatest agreement.

A third of the way through this project, the results are encouraging. In making use of the Project A LRDB to extend validity information from a relatively small sample to potentially all entry-level MOS, SynVal represents a major return on the Project A investment. Future research, building on this foundation, might include testing the transferability of the SynVal methodologies to the other Services as well as to the weapon system acquisition process.

II. COMPUTERIZED TESTING

Clessen Martin

Office of the Chief of Naval Operations

Washington, D.C.

CONTENTS

II. COMPUTERIZED TESTING

A. BACKGROUND

The Computer Adaptive Testing (CAT) Armed Services Vocational Aptitude Battery (ASVAB) Program is a Joint-Service effort to develop and deploy an automated system to replace the paper-and-pencil (P&P) version of the ASVAB used by the U.S. Military Entrance Processing Command (USMEPCOM) for enlistment testing. The CAT-ASVAB Program was initiated in 1979 by the then Assistant Secretary of Defense (ASD) for Manpower, Reserve Affairs and Logistics (MRA&L). The ASD (MRA&L) memorandum designated the Department of the Navy (DON) as Executive Agent for development of the CAT system through full-scale development. However, in 1985, under redirection from ASD (MRA&L), the CAT-ASVAB system life cycle departed from the original plan and was directed to develop and deploy an accelerated system which would use off-the-shelf equipment. This program became known as the Accelerated CAT-ASVAB Program (ACAP) and has three major objectives: (1) to develop a Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB); (2) to develop a microcomputer-based delivery system, and (3) to evaluate CAT-ASVAB as a potential replacement for the paper-and-pencil version of the battery (P&P-ASVAB).

The ASVAB is used by all the Services for initial qualification and classification decisions for enlisted applicants. The battery consists of ten tests: eight power tests and two speeded tests. There are six parallel forms to reduce the potential for compromise. Administration of the battery takes approximately three hours. The paper-and-pencil version is administered to about one million applicants each year.

The tests in the current version of ASVAB are: (1) General Science; (2) Arithmetic Reasoning; (3) Word Knowledge; (4) Paragraph Comprehension; (5) Numerical Operations; (6) Coding Speed; (7) Auto & Shop Information; (8) Mathematics Knowledge; (9) Mechanical Comprehension; and (10) Electronics Information.

B. PSYCHOMETRICS

Computerized Adaptive Testing (CAT) differs from paper-and-pencil testing in the mode of administration. In typical, conventionally-administered, paper-and-pencil tests all examinees are administered the same items, frequently in the same sequence.

The procedure for selecting items in a CAT administration is much different. At the beginning of the test, there is no information about an examinee's ability, so it is assumed to be average. Hence, an item of medium difficulty is chosen for administration. If the examinee answers correctly, the ability estimate is updated (in this case to a higher level). A second item, appropriate to this new estimated ability level, is selected for administration. If the examinee answers this second item incorrectly, the ability estimate is again updated (this time to a lower level). A third item appropriate for this new estimated ability level is selected for administration. If examinee responds with the correct answer, the ability estimate is again updated. This procedure of selecting an item, administering the item, scoring the response, and updating the ability estimate continues until some stopping rule is satisfied. In fixed-length testing, the stopping rule is satisfied when a specified number of items have been administered. In variable-length testing, items are administered until the ability estimate reaches some target level of measurement precision. Obviously, a hybrid stopping rule can be used. In summary, a CAT test is dynamically tailored to the individual examinee during the course of the test administration.

The adaptive nature of CAT makes for a very efficient utilization of test items. In a conventional test, all examinees receive all test items, regardless of ability level. This is very inefficient. For example, low-ability examinees receive many items that are too hard. Aside from the obvious waste of administration time, this procedure potentially introduces boredom and careless responses.

In an adaptive test, high-ability examinees receive only items that are relatively difficult. Average-ability examinees receive items of average difficulty, while low-ability examinees receive only relatively easy items. This focusing procedure produces a significant reduction in test administration time.

While P&P ASVAB consists of ten subtests, CAT-ASVAB consists of eleven subtests. The Auto/Shop tests consist of two separately scored tests for CAT-ASVAB. The two tests of speed are administered in a fixed-sequential way and are scored with a computerized variation of conventional methodology (guessing-corrected number of right

answers, divided by the total amount of time spent on the items). The nine other tests are power tests and are administered and scored adaptively. The model used for the adaptive tests is the three-parameter logistic model. The items were calibrated by joint maximum likelihood, with the theta measure of ability having a mean of 0 and a standard deviation of 1 in a sample of military applicants. The scoring during the test is Owen's approximation to the posterior mean; the prior distribution at the beginning of the test is normal with a mean of 0 and standard deviation of 1. Item selection during the test is based on the item information function, with constraints on the frequency with which the item is used. The score at the conclusion of the test is the posterior mode, recomputed from the original prior distribution and from the examinee's responses. Until CAT-ASVAB norms can be developed from a nationally representative sample, the scores will be transformed to the number-right metric of the P&P-ASVAB; this is accomplished by equipercentile equating, which results in the same distribution of scores as the P&P-ASVAB and, therefore, does not alter the percentage of applicants who qualify for military entrance or for jobs in the Services.

When CAT-ASVAB was being developed, item response theory was used in the study of many of its properties. Model-based simulations were used to compare the score information of CAT-ASVAB and P&P-ASVAB. Simulations and real data analyses were used to assess the effects of mis-specification of CAT-ASVAB item parameters. However, because CAT-ASVAB will initially be equated to P&P-ASVAB and used in place of P&P-ASVAB, CAT-ASVAB results are being compared with P&P-ASVAB results to ensure that the necessary properties of the traditional instrument are preserved in the CAT-ASVAB equated score metric. These properties include reliability, score-conditional variance, and invariance of equating across subgroups of examinees (e.g., women and minority groups). Invariance of equating is of some concern even though a study of the experimental CAT-ASVAB showed that different subgroups give quite similar equating functions; CAT-ASVAB is expected to be slightly more precise than P&P-ASVAB, which could make the equatings discernibly population-dependent. Another question of importance is the extent to which the equating obtained from non-operational testing is valid for operational testing; this question will be studied by comparing the distributions of equated scores from the operational and non-operational administrations of CAT-ASVAB and P&P-ASVAB.

C. PSYCHOMETRIC PRODUCTS

Since the Joint-Service CAT-ASVAB Technical Committee was formed in 1985, it has discussed 38 projects conducted prior to and during the operation of the Accelerated CAT-ASVAB Program. Twenty-five of the projects pertained to procedures for the development of the present CAT-ASVAB forms; 21 of these 25 resulted in technical recommendations which were forwarded to the Joint-Service CAT-ASVAB Working Group; most of the recommendations are summarized and documented in the CAT-ASVAB Psychometric Decision List, to be published as a technical note. Six of the 38 projects pertained to evaluating influences on the precision of CAT-ASVAB scores; results of three of these projects have been briefed to the Committee; data analysis plans for the remaining projects have also been briefed. Four of the 38 projects pertained to the equating of CAT-ASVAB and ASVAB; data analysis plans have been briefed to the Committee and are being documented in a technical report.

Documentation on the CAT-ASVAB Technical Committee's discussion of the 38 projects is contained in the Committee's minutes. That documentation and the basic research on which the 38 projects are founded support three conclusions. First, ACAP psychometric projects are careful applications of basic research and standard psychometric methodology. Second, the projects pertaining to procedures for the development of CAT-ASVAB will result in psychometrically sound CAT-ASVAB forms. Third, projects currently in progress will address questions of the psychometric feasibility of CAT-ASVAB as a substitute for ASVAB in the near term and in a nationwide implementation.

D. ACAP ACTIVITIES

The Accelerated CAT-ASVAB Project (ACAP) involves six field activities: (1) Pre-Test, (2) Medium of Administration, (3) Cross-Correlation, (4) Preliminary Operational Check, (5) Score Equating Development, and (6) Score Equating Verification.

1. Pre-Test

The Pre-Test was designed to evaluate the human-computer system interaction. The ACAP battery was administered to 231 military recruits and 73 high school students, representing the full range of cognitive ability. Some students were obtained from high school special education classes to bolster the sample of low-ability examinees.

Each of the participants took a questionnaire upon completion of the ACAP battery. This instrument was designed to obtain information on instruction comprehension, fatigue, etc. Between four and eight examinees from each testing session were selected for an in-depth interview. The selection was done to ensure adequate representation of persons finishing both rapidly and slowly.

Results were encouraging. The examinees found the CAT-ASVAB faster and easier than paper-and-pencil tests they had taken. They liked the fact that it was self-paced, and that less writing was involved than in the ordinary, paper-and-pencil test. Some persons expressed a dislike for the fact that they could not go back and change their answers after moving to another item. A few examinees indicated that their eyes became tired. The Pre-Test was completed in November 1986. Based upon information from the questionnaire and interviews, the administrative instructions were revised, reducing the reading grade level from the eighth to the sixth grade level.

2. Medium of Administration

The Medium of Administration Study is designed to evaluate the effect of the calibration medium of administration on score precision. The subjects were recruits at the Naval Recruit Training Center in San Diego. Forty-item conventional tests were constructed for General Science, Arithmetic Reasoning, Word Knowledge, and Shop Information. Persons were randomly assigned to one of three groups. The first group was administered the tests on computer. Their data were used to obtain a computer-based calibration of items. The second group took the tests in a paper-and-pencil mode. Their results were used to obtain paper-and-pencil calibration information. Each of these calibrations was used to estimate the ability of examinees assigned to the third group of examinees, who took the items on computer.

Data collection for the first phase (involving the four tests) has been completed. The number of examinees taking the battery on the computer was 1,978, while 978 examinees took the paper-and-pencil version. Analyses are currently under way. The same procedures will be followed for Paragraph Comprehension in the second phase.

3. Cross-Correlation

The Cross-Correlation Study is designed to compare the precision of CAT-ASVAB and P&P-ASVAB. The goal was to test 1,250 recruits from the Naval Recruit Training Center in San Diego. The operational P&P-ASVAB was one of the following forms: 11A,

11B, 12A, 12B 13A, or 13B. There were two forms of CAT-ASVAB (non-operational). Finally, there were two non-operational P&P-ASVAB forms employed: 9B and 10B. The operational P&P-ASVAB was the battery that the examinees had taken to enlist in the Navy. The first group took P&P-ASVAB Form 9B, then P&P-ASVAB Form 10B. In each case, the second test was administered about five weeks after the first non-operational test.

The first test phase has been completed, with 1093 examinees taking the two non-operational P&P-ASVAB forms. Subsequently, in the retest phase, 786 persons took the CAT-ASVAB, while 761 took the P&P-ASVAB. The data base for this study has been constructed and analyses have been initiated.

4. Preliminary Operational Check

The Preliminary Operational Check was designed to demonstrate the communications interface between the Accelerated CAT-ASVAB Project (ACAP) System and the U.S. Military Entrance Processing Command (USMEPCOM) System. The test took place at the Seattle Military Entrance Processing Station (MEPS).

The testing procedures were performed jointly by personnel from the Navy Personnel Research and Development Center (NPRDC) and USMEPCOM. Data from 31 examinees, tested in five different sessions, were used in the study. These data were loaded onto the Data Handling Computer at the MEPS, then transferred to the MEPS System-80 minicomputer. Comparison of the data before and after transfer showed the test was completed with perfect accuracy.

Future plans involve merging and editing ACAP results on a MEPS System-80, then telecommunicating the information to USMEPCOM Headquarters.

5. Score Equating Development

The Score Equating Development Study is designed to equate CAT-ASVAB with P&P-ASVAB. Subjects were applicants for enlistment at six MEPS and their satellite Mobile Examining Team Sites (METS). The following six MEPS/METS complexes were selected to be representative of the nation as a whole: San Diego, Richmond, Seattle, Boston, Omaha, and Jackson.

The operational measures included P&P-ASVAB Forms 10A, 10B, 11A, 11B, 13A, and 13B. There were two forms of the CAT-ASVAB (both non-operational).

Finally, P&P-ASVAB Form 8A was used as the non-operational reference battery. Subjects were assigned to one of three groups. The first group took CAT-ASVAB Form 1, then the operational P&P-ASVAB. The second group took CAT-ASVAB Form 2, then the operational P&P-ASVAB. The last group took the reference battery (P&P-ASVAB 8A), then the operational P&P-ASVAB. In each case, the testing was done on the same day, or on successive days.

Testing has been completed in all sites. The following sample sizes were obtained: San Diego (1313); Richmond (1965); Seattle (1270); Boston (1868); Omaha (914); and Jackson (1021).

Results to date have been encouraging. The microcomputer delivery system has performed satisfactorily, exhibiting fewer problems than anticipated. The logistics of administering the battery in the numerous, heterogeneous testing sites has presented no problems which have not been overcome.

6. Score Equating Verification

The Score Equating Verification Study is designed to evaluate the effect of examinee motivation upon item calibration and equating. The subjects will be applicants for military service coming through the same six MEPS/METS complexes used in the Score Equating Development Study. The measures will include two forms of CAT-ASVAB and one form of P&P-ASVAB (8A). The CAT-ASVAB scores will be based upon the score Equating Development Study. An equipercentile equating will be performed for subsequent operational use.

The planned schedule for score equating verification involves starting San Diego data collection in February 1990. Scheduled completion date for data collection in Jackson is April 1991.

E. FUTURE TESTING

New technologies are required for enhanced capability to classify personnel based on a broader profile of individual skills. This capability is urgently needed given that a critical problem is being addressed. This capability will continue to be needed not only in the coming decade but in the next century as well.

Schmidt, Hunter and Dunn (in press) showed that a test battery validity increase over the current ASVAB of 3 percent would result in the equivalent of $83 million annually

in performance improvement in the Navy. In their study of crew characteristics and ship condition, Horowitz and Sherman (1977) found that an increase of one percentage point in the average Shop Practices Test scores of Boiler Technicians on two-screw, 1200 psi ships would lower Casualty Report downtime by an average of 138 hours per month. Based on empirical results of these and other like studies, increased test validity could have a large impact, both in terms of monetary costs and in terms of fleet readiness.

The Navy's goal has been to use the unique capabilities of the computer to administer stimuli or measure behaviors that paper and pencil tests cannot handle. By so doing, the hope is to provide better measures of ability to broaden the scope of abilities measured.

Work on cognitive speed has been motivated by the work of Nettlebeck, Jensen and others showing that there were substantial correlations between measures of general intelligence and measures of speed in elementary cognitive processes. The approach was to explore several variations on ways of measuring cognitive speed, in order to boost the correlation with g while keeping the test virtually knowledge-free.

In the Reaction-Time Paradigm, the computer presents a stimulus or decision to be made, and the latency of the examinee's response is measured. Results showed adequate retest reliabilities and incremental validities for two-choice reaction time tests. One variation, the Arrows test, was sufficiently complex as to boost the correlation with general ability.

In the Inspection Time Paradigm, stimuli are flashed on the screen very quickly, and accuracy of response, not latency, is the important measure. Adequate reliabilities and some incremental validities were obtained with such tests, but these tests cannot be recommended for operational testing because results are sensitive to variations in illumination and to eyesight.

In the Machine-Paced Paradigm, the computer presents information to be processed at rates that strain the examinee's ability to handle the flow of data. The Mental Counters test was constructed to be such a cognitive speed measure, but it is also a working memory task.

Another kind of ability that the current ASVAB fails to measure adequately is Spatial Ability. Paper-and-pencil tests of spatial abilities have been around for nearly 90 years, and have shown good validities with school and job performance in many areas. But computerized measures have several advantages over paper-and-pencil tests. By

presenting problems in stages, the computerized test can require a clarity and stability of imagery that may not be needed when a whole item is displayed at one time. The computer can also measure latency during the several steps in answering an item.

Thus, measures of the information processing components that contribute to the examinee's performance can be obtained. The Integrating Details test is one such spatial ability test, where the item pool has been constructed to systematically vary the features that contribute to item difficulty.

Computers can be used to get better estimates of ability by measuring processes, not just final performance. This is being done with the spatial ability tests mentioned above, and with tests now under development to measure reasoning processes with verbal and figural rule induction problems.

NPRDC has carried out studies on 30 different samples so far. Results from the Integrating Details and Mental Counters tests will be presented briefly. NPRDC contracted with Dr. Earl Hunt and Dr. James Pellegrino to develop a computerized battery of six static spatial ability tests and five "dynamic" (animated) tests. These tests were programmed on Apple II computers and administered to 170 college students along with eight reference paper-and-pencil tests.

A factor analysis was performed with the static tasks. The first two factors can be identified as latency and accuracy. The Integrating Details test has very high loadings on both factors, which is why it was selected for further study on Navy recruits.

Each item of Integrating Details consists of a disheveled puzzle and completed puzzle. If the puzzle elements are correctly fused the resulting object will either match or not match the completed puzzle. However, the puzzle pieces and the completed object are not presented simultaneously. Instead, the subject is first provided the puzzle pieces. The puzzle pieces have lettered edges (A, B, etc.) showing how they connect with the other elements. The subject's task then, is to connect the elements and form a unitary object. This object must then be remembered. When the subject completes this portion of the item, he/she presses a key. Following the keypress, the puzzle pieces are removed and a completed object is presented. The subject must then decide if the completed object matches the object created from the puzzle pieces. If the two do not match the examinee responds "Different;" if they do match, a "Same" response is required. For each item two latency measures and one accuracy measure are recorded. The first latency is the time the subject viewed the first screen (containing the puzzle elements), called "Presentation

Latency." The second latency is the time spent viewing the second screen (containing a completed object), called "Decision Latency." Accuracy is simply whether the examinee was correct or incorrect.

The test measures complex spatial/figural problem solving. The accuracy score is correlated with the Spatial subtest from the Differential Aptitude Test and Raven's Progressive Matrices. It is uncorrelated with very simple tests of figural problem solving such as perceptual speed tests and tests of two-dimensional rotation as well as tests of verbal and numerical abilities. The latency scores are virtually uncorrelated with any accuracy based measure of any ability.

In the early development stages coachability was investigated, which led to the constraints in creating mismatch trials. There are no known easily learned strategies that can generally improve performance. If the test is administered twice, a month apart, there is, on the average, a gain of 0.43 standard deviation units for the latency based measures, but no gain in the accuracy scores. There are no known subgroup differences. Importantly, the college sample did not show any sex differences. The Navy samples contained no women. The high school sample showed a correlation of 0.17 between accuracy and gender (males are favored), but no correlation between the two latency measures and gender. The high school sample was less than 10 percent minority so no analyses were performed on race. The latency measures have split half reliabilities from 0.65 to 0.68. The accuracy scores have split half reliabilities of 0.73 to 0.79 and retest-reliability of 0.70.

The Decision Latency measure correlated 0.36 with an overall hands-on job performance measure in the Navy Machinist Mate sample. The correlation between accuracy and high school rank was 0.42 (N = 299) and 0.24 between high school rank and Presentation time. In the Machinist Mate sample, the Decision Latency measure accounted for 15.9 percent of the variance in the hands-on-performance measure after removing the Machinist Mate selection composite (of the ASVAB). AFQT alone correlated 0.33 with the criterion. When Decision Latency was added to the regression, the multiple correlation was 0.51, an increase of 139 percent in criterion variance accounted for.

A detailed information processing performance model has been constructed for the test and validated against eight separate predictions. For example, the performance model for presentation latency states that individuals encode an element, search the array for the matching letter, then synthesize the two pieces at the correct location and store the product

in memory. Therefore, Presentation time should be a function of the number of iterations of the encode-search-synthesize sequence. Across the items, the correlation between the number of elements and Presentation time was 0.93 (N points = 60). Similarly, the model has generated predictions relating decision time and accuracy to number of pieces, shape complexity, and answer choice similarity. Collectively, these data strongly support the construct validity of the test in terms of the hypothesized model for problem solution and item difficulty.

The Mental Counters test is the outgrowth of work on Cognitive Speed. It is the first test to use the Machine-Paced paradigm. The earlier Reaction Time paradigm is subject to speed-accuracy tradeoffs and the Inspection Time paradigm is subject to variations in illumination, but the Machine-Paced paradigm obviates all these difficulties. As stated earlier, the goal of the Cognitive Speed effort has been to find knowledge-free measures of g. The Counters test, as we shall see, has high construct validity as a measure of g, as compared with the earlier cognitive speed tests, and is relatively knowledge-free.

In the Mental Counters test, subjects must keep track of the values of three independent "counters." The values change rapidly and in random order. The difficulty of the task comes from having to simultaneously hold the three counter values in memory and make a series of rapid counter updates based on simple arithmetic calculations. If counter adjustments are performed too slowly, the updates themselves become additional data that must be held in memory, even as new updates are required. Individuals who cannot make rapid counter adjustments will eventually experience a performance "breakdown" as capacity is exceeded.

The test was designed to measure g. It was based on the view held by some scientists that g reflects the fundamental speed/efficiency of mental (especially working memory) processes. The Mental Counters test does seem to measure a generic aptitude underlying various complex problem solving tasks, since it is well correlated with problems requiring logical reasoning, spatial visualization, mathematical aptitude, and mechanical comprehension. Some practice effects have been observed, but post-test interviews with subjects have revealed no specific techniques or strategies that might be coached. The split-half reliability is 0.93, and the test-retest reliability was 0.64 in one sample and over 0.70 in another sample. Counters was significantly correlated with high school GPA ($r = 0.40$, $p < 0.01$) in a study done on university students. Concurrent validity study on Navy Electronics Technicians is in progress. Larson and Succuzzo

(1987) showed that Counters had high correlations with the Raven, spatial ability, and SAT scores. Comparison with the ASVAB showed that Mental Counters reliably measures ability not measured by the ASVAB.

F. COSTS

A major report on the cost/benefit analysis of CAT-ASVAB was published in March, 1988. The purpose of the report was to assess the cost/benefit implications of implementing CAT-ASVAB in the Military Entrance Processing Command (MEPCOM) under four alternative concepts of operations. The four concepts and the current paper-and-pencil testing alternative were compared in terms of life cycle costs in three major areas: R&D, Equipment Investment, and Operations and Support. The economic benefits of CAT-ASVAB were estimated by application of the Personnel Utility Formula (Cronbach and Gleser, 1965). The four alternative concepts range from a centralized and a MEPS only testing facility to a fleet of 283 mobile testing vehicles and 50 more high volume testing sites in addition to the present 70 MEPS. Intermediate concepts involved increasing the number of high volume sites to 273 plus the existing 70 MEPS. Another alternative considered a Computerized Adaptive Screening Test (CAST) which would be administered by recruiters and complete CAT-ASVAB testing would then take place only at the existing MEPS.

Crucial in the computation of the dollar benefits is the assumed increase in predictive validity of any alternative system. In computing the utility benefits associated with the CAT-ASVAB system, the contractor based the increase in predictive validity on improved reliability. Based on simulation studies, this could result in a .002 increase in predictive validity (Segall, in preparation). However, results showed that all of the CAT-ASVAB concepts of operation had higher total life cycle costs than the paper-and-pencil ASVAB in a relatively inexpensive test (approximately $15,000,000 annually for 1,000,000 tests) and that the investment necessary to replace it with a computerized test is large ($15 to $40 million). Consequently, it was concluded that significant increases in validity were needed to make any new computerized enlistment testing cost-effective.

Beginning in 1989, the CAT-ASVAB program was redirected in order to accelerate the validation of new types of computerized tests. All of the Armed Services are pursuing personnel testing research programs which focus on improved measurement of human abilities through the use of computers. At the time of this writing, a Joint-Service technical advisory test selection panel has been appointed to review the range of computerized tests

which have been developed within each of the Services' personnel R&D programs. The plan is to select a computerized test battery which is more likely to improve upon the validity of the present ASVAB. Once the Joint-Service computerized test battery has been chosen, it will be programmed on the present CAT-ASVAB hardware system and validated against both training and job performance criteria. The results of the validation analyses will provide the basis for a revised cost-benefit analyses of an enhanced CAT-ASVAB system. It is expected that the net utility of adding new types of computerized tests to the ASVAB will more than offset the costs associated with implementation of a DoD computerized aptitude battery throughout the nation.

REFERENCES--CHAPTER II

Larson, G.E. and Saccuzzo, D.P. (1987). *Analysis of test-retest reliability for a battery of cognitive speed tests* (NPRDC TN 988-10). San Diego, CA: Navy Personnel Research and Development Center.

Horowitz, S. and Sherman, A. (1977). *Crew characteristics and ship conditions* (CNS 1090). Alexandria, VA: Center for Naval Analyses.

Hunt, E., Pellegrino, J.W., Abate, R., Alderton, D.L., Farr, S.A., Frick, R.W. and McDonald, T.P. (1987). *Computer-controlled testing of visual-spatial ability* (TR 87-31). San Diego, CA: Navy Personnel Research and Development Center.

Schmidt, F., Hunter, J. and Dunn, W. (in press). *Potential utility increases from adding new tests to the Armed Services Vocational Aptitude Battery (ASVAB).* San Diego, CA: Navy Personnel Research and Development Center.

Segall, D. (in preparation). *Economic benefits of the CAT-ASVAB: Assessing effects of measurement errors.* San Diego, CA: Navy Personnel Research and Development Center.

III. USE OF APTITUDE TESTS IN MILITARY SELECTION AND CLASSIFICATION

William Alley

Air Force Human Resources Laboratory

Brooks Air Force Base, Texas

CONTENTS

III. USE OF APTITUDE TESTS IN MILITARY SELECTION AND CLASSIFICATION

A. INTRODUCTION

In the 100 years since the term mental test was first introduced,* the science and technology of mental testing for employment purposes has expanded enormously. What has been learned in this period and what still remains to be explored has been subject of much recent discussion and thought both inside and outside the military. This paper will attempt to provide a brief overview of this domain with special emphasis on the use of tests for selection and classification of military personnel. Three major streams of R&D activity are highlighted--each of which is critical to cost-effective use of tests in this context. The first centers on the tests themselves: basic concepts and content, assumption/premises and measurement strategies. The second focuses on the job and deals with the process of how one determines directly or indirectly what mental abilities are required. Findings from this area determine in large measure what abilities need to be tested to ensure effective performance. The third and final stream involves the fusions of both job and test information in the actual decisionmaking framework that determines who should be selected and into what job categories those selected should be classified.

B. WHAT WE KNOW

1. Test/Measures

One of the most consistent findings in the testing literature has been that individual differences in aptitudes exist and can be reliably measured in a relatively brief period of time. We also know that these differences are stable enough over time to be of value in estimating future performance potential. By far the most abundant source of such findings comes from evidence accumulated over literally hundreds of validity studies that have used initial training outcomes, job performance and tenure as criteria. Not only do higher aptitude people learn faster but they achieve higher levels of skill acquisition over the

---

* By J. M. Cattell in 1890.

course of instruction (i.e., achieve higher grades). There is also some evidence that they retain information longer and are quicker to relearn it. Recent efforts to extend the frame of reference to the on-the-job performance domain has shown similar albeit more modest relationships. People with higher aptitudes perform better on the job than contemporaries with lower aptitudes. Finally, higher aptitude people are more likely to finish their obligated service successfully than to attrite early.

What kinds of aptitude measures are most useful in this regard? Most large testing programs use what are termed multiple aptitude batteries. The Joint Service Armed Services Vocational Aptitude Battery (ASVAB) is a typical exemplar. Based on factor analytic results, we know that aptitude batteries measure both verbal and quantitative skills, any number of more specific abilities, e.g., clerical speed and accuracy, as well as various domains of technical knowledge, e.g., mechanics.

Although the ASVAB has multiple subtests, it is nonetheless considered to be heavily "g" loaded. That is, much but not all of the variance in individual tests scores can be accounted for by a single general factor hypothesized to underlie all cognitive test batteries. The issue of how much differential value the battery contains (for assignment purposes) is still an unresolved issue. All of the Services construct several composite measures from different combinations of subtests--one each to correspond with a like-named job cluster. The Air Force has four composites/job clusters: Mechanical, Administrative, General and Electronics. Thus, the Services operate as if more than one measure is necessary to account for meaningful differences between people--and correspondingly, for meaningful differences in jobs.

Another rather prominent feature of most modern aptitude tests is the pattern of test scores exhibited by gender and ethnic subgroups. Males for the most part show elevated scores on tests and composites that tap technical knowledge domains. Women, as a group, show higher aptitude in administrative areas while both groups score about equally well in the general ability domain (high verbal skills for females is almost exactly offset by higher numerical facility in males). Ethnic differences follow a general pattern of lower scores for minority group members (Blacks and Hispanics) on the order of 1/2 to one standard deviation as compared to the majority group. This creates a certain amount of controversy with regard to "fairness" in testing. Most psychometricians distinguish, however, between subgroup differences on tests and situations where there is likely to be systematic over or underprediction in relation to a performance criterion. The latter is considered to be evidence of unfair testing practices. Most empirical investigations find that where evidence

of bias is found to exist--a relatively uncommon finding--the tests typically overpredict for minority members rather than underpredict. The magnitude of the validity coefficients computed within subgroups is generally always about the same.

As with most of the aptitude batteries used for employment purposes, there is an explicit assumption that applicants taking the test are doing the best they can without benefit of artificial aids (compromise) nor are they intentionally scoring below what they would have normally (malingering). Both types of inappropriate test behaviors are of considerable interest to the Military Services. At this point, techniques for detecting compromise/malingering are fairly crude in the sense that they are used primarily in the aggregate for detecting trends over time. Some Services use them occasionally to channel selected groups or individuals for retesting. But due to the large "false positive" detection rate, the procedures have not gained widespread institutional use.

A particularly promising testing technology has emerged during the past 15 years called Item Response Theory (IRT). This is an extension of classical test theory which focused on the test as the primary unit of analysis. In IRT, each item in a test is hypothesized to yield valuable information about the attribute to be measured. The information is embodied in an item characteristic curve that can be expressed (depending on the assumptions) as a one, two or three parameter model where the parameters correspond to (a) a basal value--correction for guessing (b) an item difficulty parameter and (c) an item discrimination parameter.

If the items can be administered by computer, IRT provides a capability to make a test adaptive in the sense that the selection of the second and subsequent items to administer can be made to depend on responses to earlier items. The test can then be made to converge more quickly and some say more accurately on an estimated ability level. The full potential of these techniques has not been fully exploited. In fact, it is the subject of a fairly large-scale R&D effort at present (CAT-ASVAB).

Of even more importance, however, are the implications of computer administration to the search for new cognitive predictors of training and job performance. It is generally recognized that certain attributes are poorly measured by paper-and-pencil tests if they can be measured at all. These would include speeded response, attentional resource allocation on dual tasking, perceptual/psychomotor abilities, among others. Each of the Services has programs of this sort--some theory based, i.e., Learning Abilities Measurement Program (LAMP) and some more empirical and programatically-oriented (Project A).

Until this point, discussion has centered on the more traditional cognitive tests (verbal, quantitative, etc.) currently in the battery. There is general agreement that other types of assessment could provide valuable adjuncts to aptitude measurement. Among these are interest inventories and bio-data forms. Interest inventories have a long-standing history of usage in civilian guidance programs. There has been recent evidence that they would contribute over and above the ASVAB in predicting subsequent satisfaction with assignment and longevity in the service. Applications to date have so far been limited. The Air Force, for example, administers one to all incoming recruits but only uses results for those entering without a guaranteed job--approximately 50 percent of the total.

Bio-data forms have also been shown to have significant relationships with propensity for early attrition. Both the bio-data and vocational interest inventories are being explored at the Service and DoD levels. The savings to the government, if even a small proportion of the early attrites can be eliminated through interest assessment or bio-data forms, is enormous--on the order of tens of millions of dollars.

The whole issue of expanding the "criterion space" remains an issue unto itself. Previously mentioned was a major Joint Service effort to construct on-the-job performance measures. But in fact, selection and classification decisions have potential effects beyond predicting how well a person might perform in training or on the job. We also know that personnel tenure can be altered, sometimes dramatically, by including different predictors in the selection system or by altering entry level requirements. Evidence has shown that higher aptitude personnel have a greater propensity to complete initial obligated service commitments. On the other hand, they do not reenlist in as high a proportion since there are many more civilian job opportunities available to them. We have only begun to understand how aptitude affects performance and tenure and how total productive capacity--defined as a joint function of performance and retainability--can be maximized through more carefully designed personnel selection systems. This would apply to both first and subsequent tours--the latter of which are not well understood.

Finally, we know that the whole process of test development, going from items to finished tests, is still cumbersome, time consuming and prone to error. Computer-assisted test development is in its infancy and could save valuable resources while at the same time producing a more robust product.

2. Jobs

As testing technology continues to evolve, a parallel issue concerns what the jobs actually entail, how they change over time, and what these changes imply for test development efforts. By concentrating on the many ways that people differ in their aptitudes, it is easy to lose sight of the fact that Selection and Classification systems are designed to support a particular job structure--and if the requirements of jobs are not well understood, then the employer runs the risk of selecting on the wrong dimensions.

The ideal process by which job requirements are determined (in a data-rich environment) is by conducting longitudinal validity studies in which people are measured prior to entry along dimensions hypothesized to be relevant to training and/or job success. Follow-up measures of performance, however defined, are statistically evaluated in relation to the selection measures. Requirements are inferred based on what subsets of predictors relate and their relative contributions to the equation. This is obviously the most direct method but it suffers from some serious practical constraints. First, the only requirements that can be identified are those embodied a priori in the test measures. Second, due to the relative abundance of data on training outcomes vice later performance, the requirements identified are those most closely associated with training and may only reflect part of the total job requirement. Third, such studies are often impractical to conduct due to time considerations, low throughput, or difficulty in obtaining sufficient follow-up measures.

A vast literature in job analysis provides a number of (as yet unexplored) alternatives to validity stu s--alternatives that offer some insights that even the classical approach can seldom provide in a timely way (i.e., implications of rapid job changes and projection of future requirements).

We know that jobs can be measured (or categorized) on the basis of the human abilities they require, on the basis of tasks to be performed or on some hybrid involving both person and task-oriented approaches. From findings in this area, we know that some jobs differ in their requirements for mental ability as compared to other jobs and in comparisons with their own requirements over time. Along what primary dimensions do they differ? The issue has not been completely resolved.

However, it is generally recognized that jobs differ in terms of (a) the amount of time required to learn to perform them successfully, sometimes called learning difficulty or complexity; (b) the degree to which they require numerical and/or verbal facility; (c) requirements for speeded response; (d) environmental pressures associated with

acceptable performance, i.e., stress producing; (e) extent to which the incumbent deals with physical as opposed to conceptual problem solving; and (f) extent to requirements for prior knowledge/experience in a particular technical domain and others of perhaps lesser consequence from a mental abilities standpoint.

The Services for the most part view these differences to be substantial enough to warrant the use of separate composites for classification purposes. The test composites and their associated minimum cutoffs for specific jobs are constructed on the basis of empirical data (validity studies), job analysis results and expert judgment. This would suggest that the military derives value from the multiple composite approach. The issue has not been completely settled as mentioned earlier. There are those who would argue that it may be sufficient to distinguish jobs solely on the basis of overall aptitude ("g" requirements). In any event, our understanding of occupational requirements can be better informed by job analysis, particularly those that proceed from detailed task lists. The Air Force has experimented with a fairly complex system involving the use of "task difficulty benchmark scales" which when applied and aggregated to the job level, provide valuable information about the relative difficulty of jobs.

The concept of learning difficulty is important because it is the analog of individual aptitudes. Thus, while people can be ordered along one or more continua based on their ability to learn, so can task and jobs be ordered along a parallel continuum based on their difficulty to master--usually expressed as time to learn.

3. Person-Job Match

One of the more unique developments in the selection and classification domain concerns the general problem of information fusion at the level of decisionmaking about who should be assigned to what job. Given an array of jobs to be filled and a list of applicants to fill them, how can information about the characteristics of people be combined with the attributes of the jobs to make an appropriate match--particularly when there are multiple, competing, and often incommensurate objectives that have to be considered?

How can one balance efficiency (making sure the jobs are filled in a timely manner) with effectiveness (making sure the best candidates fill each job)? Tradeoffs must be made between delaying a possible fill action in the hopes of finding a better match in an uncertain future. In response to this problem, each Service has developed (either experimental or operational) assignment systems modeled after a novel conceptualization of the problem wherein N people must be assigned to J jobs such that the sum of the individual payoffs

across the whole system yields a maximum value. Payoffs to the Service are determined by a procedure called policy specification--a process by which a policymaker or panel constructs a mathematical representation of a policy decision based on multiple input variables.

Input variables can include the full array of personnel (aptitude, preferences, etc.), job (priority, learning difficulty) and management factors that impinge on the decision. The most unique feature of the policy-derived systems is their flexibility. Decisions can be made on the basis of full, partial or even a complete absence of empirical data. Yet, the mathematical representation is replicable and can be made to evolve over time as more is learned about the tradeoffs. Through experimentation and use, these systems have been shown to operate in a way that yields total payoff values in excess of what can be achieved with manual processing with corresponding benefits in the component variables and functions. Both the Army and the Air Force have demonstrated that optimizing the payoffs in either a batch or sequential process can reduce casual time, improve the numbers of people assigned to their first job choice, increase projected training performance, provide for higher potential job satisfaction and reduce premature attrition. With the large number of entry-level personnel, documented savings of these systems is in the 100 million dollars/year range.

C. WHERE WE SHOULD BE GOING

1. New Tests/Measures/Assessment Strategies

The search for new ways to measure individual differences that might have implications for military performance should be concentrated in four broad areas: (a) cognitive, (b) spatial, (c) perceptual-psychomotor and (d) other non-cognitive measures.

In the cognitive domain, we need to be concerned with extending the ASVAB into areas that presently are not measured or not measured very well. R&D activities need to proceed along two fronts, each of which benefits from the other: (a) basic theory and model building coupled with (b) empirical investigations of what is actually related (beyond ASVAB) to eventual performance. Areas showing the most promise take advantage of recent advances in the field of cognitive psychology. In place of the "fixed abilities" (i.e. "g," or verbal and quantitative abilities), we should be looking toward more fundamental component progresses--e.g., speed and accuracy of cognitive processing,

capacities for short (working) and long term memory and retention, prior declarative and procedural knowledge, etc. The most innovative approaches will capitalize on the added precision and storage capacity of the microcomputer as a testing device. Unlike paper-and-pencil tests, administration via microcomputer brings not only additional control to the test situation but also extends the domain of what can be measured effectively (i.e., response latency, multiple tasking, audio/visual stimulus cues and multiple response modes).

The ASVAB is especially weak in the spatial and perceptual-psychomotor domains. Here again, there is an opportunity to exploit the enormous potential of the microcomputer to present spatial stimuli that are impossible to replicate with paper-and-pencil tests. The same holds true for extending our selection measures to include perceptual-psychomotor abilities that have shown special promise in situations where performance demands are especially high. These would include such jobs as air traffic controller and fighter pilot.

Finally, the whole area of non-cognitive measures--(bio-data forms, vocational interest inventories, security assessments, etc.) remains untapped. The research community has come to realize that selection procedures should not focus exclusively on increasing expected competencies--but instead should look at a broader perspective that considers the length of expected service (attrition/retention) and performance qualities that are *not well predicted with cognitive test batteries*--leadership/management, integrity and organizational commitment.

Traditional reliance on training performance measures as the sole criterion reference should decrease in the future. For the past seven years, each of the Services has mounted large efforts to construct on-the-job performance measures (hands-on as well as other alternative procedures). There is a research data base available at present that is unprecedented in the research literature. It needs to be probed to determine (a) how present tests relate to job criteria; (b) how new tests add to the prediction of job criteria and (c) how alternative and potentially less expensive surrogate measures can be used in lieu of the more expensive hands-on measures.

As the next major step in this enterprise, research needs to concentrate on the effect of individual performance on crew, group and unit success. Unless it can be shown definitively how individual differences in ability affect sortie generation capacity or tank crew effectiveness, then we will have missed an important opportunity to show how entry standards influence combat effectiveness.

2. New Job/Task Analytic Procedures

Progress on this front will be made to the extent that traditional job analysis techniques, which in the military are vertically oriented toward considering requirements of single specialties in isolation, can be made to look and operate across specialties. One of the key factors, it turns out, that determines how many people are needed to perform a given amount of work, how much aptitude they must possess and how much they need to be trained is the specialty structure defined by the personnel system. Considering that some new systems technology is revolutionary rather than evolutionary, the MPT needs of the future are very ill-defined at present. Research in job and task analytic procedure is needed to better understand present and future requirements for those who maintain and operate our weapons systems.

What is needed is a structured approach to job analysis that permits analysis *across* specialties. Such a system could better inform our attempts to structure specialties based on common entry and training requirements. It would also allow a more flexible response to changes in operational requirements (i.e., central versus dispersed basing), permit more economical approaches to training for "common skills" and eventually provide the basis for estimating requirements for jobs that do not yet exist but which can be described in some detail.

3. New Selection and Classification Systems/Specifications

Future research on selection and classification systems can best be described in the context of the present multitiered system (see Fig. 1) in which at least four different levels of processing are distinguished: Level I, military (non-service specific) enlistment qualifications; Level II, service-specific enlistment qualifications; Level III, job/specialty qualifications; and Level IV, person-job match optimization.

Levels I, II, and III operate on a "qualified--not qualified" basis by specifying minimum tests scores that must be attained for entry. Level IV (P-J-M) processing assumes that all previous minimum qualifications have been met and proceeds to assign people to jobs such that the total system payoff is maximized.

Level I	MILITARY ENLISTMENT QUALIFICATIONS (NON-SERVICE SPECIFIC)
Level II	SERVICE-SPECIFIC ENLISTMENT STANDARDS
Level III	JOB/SPECIALTY QUALIFICATIONS
Level IV	PERSON-JOB MATCH OPTIMIZATION

Figure 1. Multitiered Selection and Classification System

The technology available for setting minimum qualifying standards is fairly rudimentary at present. Basically, the problem is one of settling on a cutoff value high enough to ensure training and job success but not so high as to unduly restrict the supply of applicants. This is usually done on the basis of expert judgment and consensus after relevant empirical findings have been reviewed. These might include relationships between various test scores and training success, current training attrition rates, the occupational learning difficulty of the job, and the recruiting environment, among others.

"Analytical" approaches to setting minimum standards as exemplified by the RAND cost/performance trade-off model or the Air Force's Time-To-Proficiency (TTP) model have recently been developed but these, as a general class, suffer from being either overly simplistic or much too narrow in scope to provide meaningful results. At present, the analytical approach represents more of an idealized future target than a near-term solution to the problem. Work needs to proceed on these models to make them more effective representations of real-world situations. Procedures need to be developed for measuring productive capacity in individuals and in functional units (work centers, squadrons). Simulation capabilities based on analytical models should be extended from the present focus on first-term capacities to second and subsequent terms to account for the fact that over time, the value of individual workers proceeds from primarily technical activities in the early part of one's career to primarily supervisory/management activities in the latter part.

Alternative "eclectic" approaches--that have already been adopted by the Services--attempt to weigh judgmentally the myriad of personnel, job and "management" factors that need to be brought to bear on decisions about who should be selected/classified into what

jobs. From the testing side, these approaches require tests and composites that have demonstrated validity and the ability to differentiate between alternative assignments. From the job requirements side, they require that we be able to characterize jobs in terms of learning difficulty and in terms of the relevant aptitudes necessary for success. In neither of these areas are we anywhere near an optimal specification. The job clusters presently in use by the Services (i.e., Mechanical, General, Administrative and Electronics) evolved over a number of years and are in serious need of restructuring. The same is true of the associated test composites upon which entry into the jobs within clusters is based. And within clusters, the systems employed for determining minimum cut scores are far from rigorous.

These systems also operate under the presumption that appropriate cognizance has been given to balancing the needs for efficient job fill with those of providing for effective job fill. The optimization (Level IV) systems in place today (i.e., PROMIS, PACE, EPAMs, etc.) are strong on "conceptual development" but weak in terms of follow-on "evaluation" or trade-off capability. That is, we need better techniques to track the overall payoffs over time, to decompose the total payoff into component benefits and to determine in advance the effects of alternative reconfigurations on the system prior to actually making changes.

4. Technology Base Efforts

Beyond the traditional triad of tests, jobs and P-J-M systems, there are a number of topics requiring future R&D that do not fit conveniently into the above notational scheme. These are better characterized as tech-base issues.

a. Multiple Criteria

In many real-world situations there are multiple criteria around which a selection system should be designed. These may be simultaneous, as in selection, to maximize training performance, as measured by academic grade and instructor evaluations, or they could be sequential--selection to improve training outcomes, job performance and retention. Relationships between the predictors and criteria can be mutually complementary, mutually antagonistic or in fact be independent. Similarly, some management preference usually exists for satisfying one or another of the criteria at the possible expense of the others. Research is needed to develop analytical strategies for coping with this situation that may involve both empirical and judgmental components.

b. Test Equity and Bias

Current methods for detecting and removing bias from formal selection systems are in many ways inadequate. R&D is needed to explore the most effective and efficient means for determining the presence or absence of bias and for taking positive steps to reduce or eliminate it.

c. Acquisition and Decay of Abilities

Developmental theories relating the acquisition and decay of test-relevant abilities over the age groups 18-55 years has not kept pace with the Services' need to understand this important phenomena. As a result, educational effects have not been adequately considered in developing test norms, policies for retesting are little more than ad hoc and we have little understanding how overall capabilities to acquire knowledge or perform effectively change over time.

d. Automated Test Development

New approaches are needed to streamline the cumbersome and time-consuming task of developing follow-on test forms. Automated item generation is in its infancy, and much more could be done to develop expert systems in this area. Once items are banked, procedures would be needed to extract items for parallel forms such that content, item difficulty, item discrimination, etc., were matched to the reference forms.

e. Detection of Compromise/Malingering

With large financial incentives at stake, there is much to be gained by employing possibly illegal means to increase test scores. Detection systems at present are useful mainly in the aggregate, i.e., identifying problems among groups of examinees in a particular geographic region. Additional work in this area could provide a much more accurate estimate of whether cheating has occurred with regard to specific individuals and by how much.

f. Population Changes in Ability Levels/Patterns

Currently, there are no established procedures for tracking changes in aptitudes/ abilities among the eligible applicant population. Methods and procedures are needed to chart and project changes as they occur and as they might be expected to occur over the

next 10-20 years. Such data could be very relevant to military planners who must trade off technological solutions with the demands created for the future operators and maintainers.

IV. CURRENT RESEARCH AND DEVELOPMENT ON SELECTION AND CLASSIFICATION

Bruce Bloxom

Defense Manpower Data Center

Monterey, California

CONTENTS

IV. CURRENT RESEARCH AND DEVELOPMENT ON SELECTION AND CLASSIFICATION

A. INTRODUCTION

Areas of selection and classification research and development exemplified by the work of the Army and the Navy, described above, can be placed into five broad categories: pre-enlistment screening, enlistment testing, reenlistment testing, validation criterion development, and evaluation of the impact of implementation. These areas are not mutually exclusive and are often addressed simultaneously, e.g., the Army's work on validating new kinds of enlistment tests against new types of job criteria in its Project A. Therefore, the use of this classification should be viewed as being for expository purposes only and should not be construed as a suggestion that each project falls into one and only one category.

B. PRE-ENLISTMENT TESTING

An excellent example of pre-enlistment testing research and development is found in the Army's work on the Computerized Adaptive Screening Test (CAST). This project illustrates how computer technology can be used to implement contemporary psychometric developments to (a) provide rapid service to the recruiting commands and (b) help project the image of a modern military. It also illustrates that, once the technology is in place, it can be readily modified to incorporate desired improvements. Finally, it provides an excellent example of how Joint Service efforts, in this case with the Navy, can expedite the early stages of development.

C. ENLISTMENT TESTING

Research and development of enlistment testing is the traditional focus of much of the work on selection and classification. Therefore, it is not surprising that most of the work reported here by the Navy and the Army is of this type. Efforts here can be seen in three areas: computerized testing, paper-and-pencil testing on constructs not presently

assessed by the Armed Services Vocational Aptitude Battery (ASVAB), and alternative ways of defining and utilizing composites for classification into military jobs.

Computerized administration of enlistment tests is undergoing development on two fronts. The first is the administration of the ASVAB in an adaptive form, exemplified by the Navy's Accelerated CAT-ASVAB Project. The second is the administration of non-ASVAB tests in a conventional (non-adaptive) form, utilizing the computer to assess what cannot readily be measured by paper-and-pencil testing; examples of these are the Army's psychomotor and perceptual tests and the Navy's tests of cognitive speed and complex spatial ability.

Conventional, paper-and-pencil testing of non-ASVAB constructs is being developed for a wide variety of constructs. From the Army's Project A alone, numerous measures of spatial ability, temperament and vocational interests are being developed and implemented.

The development of new kinds of enlistment tests inevitably forces a reconsideration of how the tests will be combined into classification composites. The Army's work on synthetic validation, its development of enlistment standards based on job performance measurement (see below), and its work on the Enlisted Personnel Allocation System illustrate a variety of methodologies for potentially improving classification.

D. REENLISTMENT TESTING

Selection and classification research and development traditionally have emphasized initial enlistment decisions. However, the Army's work reminds us of the importance of reenlistment decisions for the maintenance of a career force. The use of training grades plus other measures to predict second-term job performance provides an example of how test scores, or measures based on them, might be used to improve the predictive validity of reenlistment decisions.

E. VALIDATION CRITERION DEVELOPMENT

In response to Congressional pressure as well as pressure from policymakers and researchers, selection and classification validation is now being extended beyond the prediction of grades in training schools. This is clearly illustrated by the Army's use of job simulation (e.g., the tube-launched, optically-tracked, wire-guided missile simulator), multi-faceted ratings by peers and job supervisors, hands-on-performance tests and written job knowledge tests to assess a variety of aspects of a person's job-related capabilities and

performance. The other Services have parallel efforts in connection with the Job Performance Measurement project, the results of which are summarized in an annual report (e.g., Department of Defense, 1989) produced by the Office of the Assistant Secretary of Defense (Force Management and Personnel).

F. EVALUATION OF THE IMPACT OF IMPLEMENTATION

Selection and classification research and development are valuable only if some of their results produce procedures which are implemented and provide reduced costs or improved military readiness. Assessing whether such gains can be expected is the focus of some of the Navy's work on the dollar gains of improving the validity of enlistment testing. An example of the use of such an assessment is the Navy's prediction of the cost-benefits of CAT-ASVAB; the assessment resulted in a decision to redirect the CAT-ASVAB program towards incorporating new kinds of computerized tests in an effort to obtain incremental validity.

Another important effect of implementing a new selection and classification procedure is its impact on the qualification rates of women and minorities for enlistment in the Services and for classification into jobs in the Services. Eitelberg, Laurence, Waters and Perelman (1984) and Eitelberg (1988) provide these qualification rates based on the use of current selection and classification procedures. The Army provides an example of how new composites of enlistment tests could affect the job qualification rates of minorities.

The implementation of a new selection and classification procedure is based on studies indicating that the procedure, when administered non-operationally, will improve the prediction of a relevant criterion. However, when the procedure becomes operational, its validity may not be sustained. The Army provides a good example of monitoring validity after implementing classification procedures--based on the use of one-and two-hand tracking tests--at a training command and then discontinuing the testing in a specialty--use of air defense systems--where the tests were not found to be operationally predictive of task performance. Particularly important here was the effort to understand why the procedure did not work so that similar efforts can be avoided in the future.

G. SUMMARY OBSERVATION

This discussion of recent research and development on selection and classification testing clearly attests to the diversity of areas for improving procedures used to determine

eligibility for military service and to assign men and women to military jobs. It also points to important areas for improving the methodology of selection and classification studies.

Yet another point which stands out is the programmatic breadth and depth of the effort of at least one of the Services, the Army. This leads to the question of what is necessary for a Military Service to field such an effort in research and development and have it result in procedures which are implemented. The answer to the question is not provided in the Services' contributions to this white paper, because it is not provided by the substance of the research and development. The answer lies in having strong and sustained leadership at both the technical level and at the general officer level. Strong technical leadership is needed to keep the program comprehensive, coordinated and implementation-oriented. Strong general officer leadership is needed to keep the program supported at a Service's highest levels over the extended period of time necessary for successful implementation. The Army has been fortunate to have individuals who could provide this kind of exceptional leadership.

REFERENCES--CHAPTER IV

Department of Defense (1989), *Joint-Service Efforts to Link Enlistment Standards to Job Performance: Recruit Quality and Military Readiness*, Washington, D.C.: Office of the Assistant Secretary of Defense (Force Management and Personnel).

Eitelberg, M.J., Laurence, J.H., and Waters, B.K., with Perelman, L.S. (1984), *Screening for Service: Aptitude and Education Criteria for Military Entry*, Washington, D.C.: Office of the Assistant Secretary of Defense (Manpower, Installations and Logistics).

Eitelberg, M.J. (1988), *Manpower for Military Occupations*, Washington, D.C.: Office of the Assistant Secretary of Defense (Force Management and Personnel).